
RESOLVING SPECIES PHYLOGENIES OF RECENT EVOLUTIONARY RADIATIONS¹

L. Lacey Knowles² and Yat-Hei Chan²

ABSTRACT

For recently derived species and when the time separating speciation events is short, the phylogenetic distribution of taxa in a gene tree may not accurately reflect the actual species relationships. The phylogenetic tradition of relying on gene tree and species tree synonymy is not reliable under such historical scenarios. Nevertheless, recent studies have demonstrated that accurate estimates of species relationships are possible when the method of phylogenetic inference considers not only the stochastic processes of nucleotide substitution, but also the random loss of gene lineages by genetic drift—even when there is widespread incomplete lineage sorting. This simulation study examines how the broader phylogenetic context, that is, the species tree topology and branch lengths, influences the ability to recover species relationships when taxa have undergone a recent and rapid radiation. As expected, the time since species divergence and the time between speciation events influences whether phylogenetic relationships are accurately estimated. However, the influence of the timing of divergence on the ability to recover species relationships accurately differed depending on the relative position of the taxa in the species tree. Differences in the ability to recover these relationships across multiple simulated species trees highlight the potential effects of taxon sampling on phylogenetic inference at, or near, the species boundary and under high rates of speciation. By focusing attention on the species tree, rather than on the individual gene trees as the basis for interpretations about species relationships, these results also represent a fundamental shift from the phylogenetic paradigm.

Key words: Coalescence, divergence time, genealogical discord, incomplete lineage sorting, radiation, speciation, statistical phylogeography.

Within a traditional molecular phylogenetic framework, species relationships are inferred directly from the topology of the estimated gene trees (Felsenstein, 2004). Under this paradigm, phylogenetic resolution (at the most basic level) becomes a question of how many unlinked loci are necessary to provide corroborative evidence of species relationships (e.g., Miyamoto & Fitch, 1995; Poe & Chubb, 2004; Jennings & Edwards, 2005; Rieseberg et al., 2006) or number of nucleotide sites required to estimate a gene tree (or topology from combined data across multiple loci) (e.g., Walsh et al., 1999; Rokas et al., 2003). However, when the time separating species divergence is short, genetic drift is unlikely to have time to bring loci to fixation before subsequent speciation events (Tateno et al., 1982; Tajima, 1983; Pamilo & Nei, 1988). Consequently, the estimated gene tree has limited utility because the most recent common ancestor of individuals is not likely to occur within a species lineage.

The problems that incomplete lineage sorting (i.e., the failure of gene lineages to coalesce into a common ancestor before subsequent species divergence) poses for estimating species relationships are widely recog-

nized—the genealogical histories of individual loci may not faithfully reflect the phylogenetic history because of retention and sorting of ancestral polymorphism (Avise et al., 1983; Pamilo & Nei, 1988; Takahata, 1989; Maddison, 1997; Rosenberg, 2002, 2003). To avoid misleading conclusions about species relationships, arising from either discord between a gene tree and the species tree (Maddison, 1997) or widespread incomplete lineage sorting that obscures any obvious pattern of historical relatedness (e.g., Jennings & Edwards, 2005; Maddison & Knowles, 2006; Knowles & Carstens, 2007a), methods of phylogenetic inference need to take into account the process of gene-lineage sorting (e.g., Maddison & Knowles, 2006; Liu & Pearl, 2007). These methods shift the focus away from the idiosyncrasies of individual gene trees as the basis for inferring phylogenetic relationships, to estimating the history of species divergence (i.e., the species tree) directly (e.g., Carstens & Knowles, 2007; Edwards et al., 2007).

One of the conditions where the stochastic loss of gene lineages by genetic drift and incomplete gene-lineage sorting will make gene trees an unreliable basis for inferring taxonomic relationships involves

¹Thanks to Rick Mayden, and the organizers of the symposium, for inviting me to participate in the symposium. The research was funded by the following awards to LLK: a National Science Foundation grant (DEB-0447224) and the Elizabeth Caroline Crosby Fund, National Science Foundation ADVANCE Project, University of Michigan.

²Department of Ecology and Evolutionary Biology, 1109 Geddes Ave., Museum of Zoology, Room 1089, University of Michigan, Ann Arbor, Michigan 48109-1079, U.S.A. Author for correspondence: knowlesl@umich.edu.

doi: 10.3417/2006102

the rapid diversification of species. During evolutionary radiations (e.g., Shedlock et al., 2004; Nikaido et al., 2006), speciation events occur in rapid succession. Consequently, reconstructed gene trees may not accurately reflect the history of speciation among taxa because of the persistence of ancestral polymorphism across subsequent species divergence events. In fact, the expected lack of congruence among independent genes (Hasegawa et al., 1985; Slowinski, 2001) has been used to infer that species have radiated (e.g., Jackman et al., 1999; Poe & Chubb, 2004).

Rather than commenting on the discordant genealogical histories as evidence for a radiation, here we examine whether a signal of species relationships can be extracted despite the genealogical discord that accompanies rapid speciation. We consider the effects of both the timing of species divergence and the internode distance (i.e., the timing of the preceding speciation event) on recovering phylogenetic relationships when there is widespread incomplete lineage sorting. The ability to estimate the phylogenetic relationship of a pair of sister taxa is explored over a natural range of branch lengths and topologies, rather than focusing on a specific species tree (e.g., Takahata, 1989; Rosenberg, 2002), thereby providing a broad historical context for examining how the history of species divergence during a recent radiation will affect our ability to recover species relationships.

METHODS

PROCEDURAL RATIONALE

The information for estimating phylogenetic relationships is extracted from the pattern of gene-lineage coalescence (as described in Maddison & Knowles, 2006), as opposed to synonymizing the gene trees with the species tree. We focus on historical scenarios where the hazards of incomplete lineage sorting are expected to predominate—namely, recently diverged species. Because the probability that a particular gene tree for a population model can be calculated under the coalescent process (Takahata & Nei, 1985; Rosenberg, 2003; Degnan & Salter, 2005), a species or population tree can be inferred (in principle) using a full-probabilistic model (e.g., Maddison, 1997; Degnan & Salter, 2005). In practice, Maddison and Knowles (2006) showed that the species history can be accurately inferred, even if the actual probabilities of incomplete lineage sorting are not quantified under a stochastic model. Therefore, an approach that incorporates the genetic process that results in the coalescence of gene lineages below the divergence of the species (i.e., deep coalescence) during the phylogenetic inference procedure is applied here.

COMPUTER SIMULATIONS

Species trees that include a natural spectrum of topologies and branch lengths were simulated using Mesquite version 1.1 (Maddison & Maddison, 2004). Species trees were simulated for six species to have a total time depth (summed length of branches from any terminal down to the root) of 100,000 generations according to a Yule model. With an N_e of 100,000, a total time depth of 100,000 generations (i.e., a total tree depth of $1/N$) leads to considerable incomplete lineage sorting (e.g., Fig. 1). One thousand species trees were simulated, 100 of which were selected at random as species trees in which a particular taxa pair was sister (species E and F were randomly chosen as the target sister taxa), out of the 109 species trees with E and F as each other's closest relative, and species A was the outgroup. To account for the stochasticity of the genetic process of gene lineage coalescence, the ability to recover the sister relationship of species E and F was examined over 100 replicate data sets for each species tree, in which either one locus or three loci were sampled for each of the 10 individuals in each species. Gene genealogies for each replicate were simulated by a neutral coalescence (Kingman, 1982; Hudson, 1990) through Mesquite's Neutral Coalescence module, which uses an exponential approximation to avoid fully explicit modeling of individuals. A biologically reasonable effective population size (N_e) of 100,000 was used for all simulations. Population size of the ancestral species lineage was also set to the common size of 100,000; this represents a reasonable null model of speciation. Otherwise, varying the effective population size of the ancestor would imply some sort of demographic event (e.g., a bottleneck or expansion if its size was set to smaller or larger values, respectively, compared to the descendent taxa).

A species tree was then estimated for each replicate from the simulated gene trees by minimizing the number of deep coalescences between the gene trees and the species tree (for details, see Maddison & Knowles, 2006). This method is based on searching for the species trees that minimize the implied number of deep coalescences in the contained gene trees (Maddison, 1997). The number of deep coalescences was counted assuming the reconstructed gene trees were unrooted and using an "as is" taxon addition sequence, followed by subtree pruning regrafting branch swapping, saving only a single tree at any stage (MAXTREES = 1).

STATISTICAL ANALYSES AND SUMMARY OF RESULTS

To evaluate the ability to recover the sister relationship of E and F, each replicate was examined

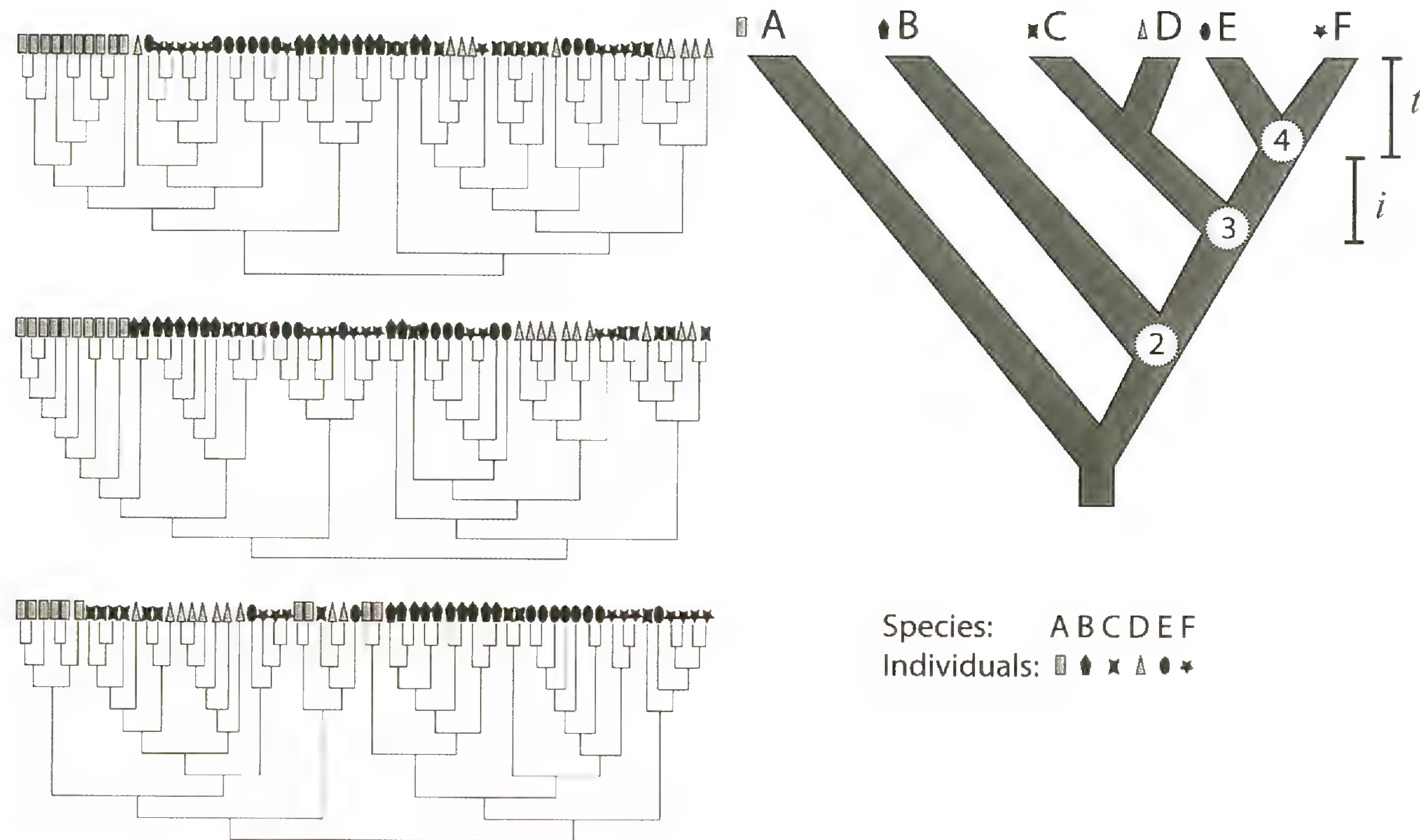


Figure 1. Example of the degree of incomplete lineage sorting apparent in a gene tree of any single sampled locus and the degree of discord among sampled loci for the recent radiation of species studied here (on the left are three representative gene trees); the effects of the species divergence time (t), internode distance (i), and nodal position (shown in the white circles) of the sister taxa relative to the other species on recovering species relationships accurately are considered here. Gene trees were simulated with 10 gene copies (i.e., individuals) per species by neutral coalescence within a simulated species tree (shown on the right) with a total depth of species tree from root to tips = $1N$, where $N = 100,000$.

and the number of times species E and F were estimated as sister taxa was recorded. To examine what factors influence the accuracy of estimated relationships, (1) the depth of the species divergence, (2) the time of the preceding speciation event (i.e., the length of the internode defining the sister taxa E and F), as well as (3) the position of the E and F species split relative to the other four taxa in the original species tree were noted (Fig. 1).

Analyses of variance (ANOVAs) were used to test for a relationship between divergence time and internode distance (Fig. 1) on the ability to recover the sister relationships of species E and F. An analysis of covariance (ANCOVA) was used to test whether this relationship differed depending on the relative position of the target taxa in the species tree (i.e., nodal position). To separate the effect of nodal position from the influence of differences in the timing of divergence on the accurate recovery of E and F as sister taxa, this analysis was carried out on the residuals from the regression of phylogenetic recovery and divergence time.

RESULTS AND DISCUSSION

For the shallow species histories considered here, there is a very low probability of reciprocal monophyly of the species (Hudson & Coyne, 2002), and incomplete lineage sorting and gene tree discord predominate (see also Maddison & Knowles, 2006; Knowles & Carstens, 2007b). With a total tree depth of $1 N_e$, which corresponds to 100,000 years assuming one generation per year and an effective population size of 100,000 for each species, the timing of species divergence and interval between speciation events is very short (see Fig. 1). For example, the average depth of the divergence of species E and F ranged from 0.01 to $0.16 N_e$ and averaged $0.1 N_e$, or approximately 10,200 years assuming one generation per year, whereas the average internode distance, or the time of the preceding speciation event, ranged from 0.002 to $0.03 N_e$ with an average of about 2300 years (Fig. 2). The rate of speciation in the simulations (i.e., six species originating with the total tree depth of 100,000 years) corresponds to the origin of a new species about every 17,000 years. Therefore, the simulations serve as a guide to the conditions affecting phylogenetic accuracy when species have undergone an evolutionary radiation (e.g., Mendelson & Shaw, 2005). The range of species divergence times represented in the simulation study is also common to species-level studies (Arbogast et al., 2002), especially those involving Pleistocene divergences (e.g., Knowles, 2000, 2001; Masta & Maddison, 2002; Hewitt, 2004), making the study broadly relevant to

the general difficulties with reconstructing species relationships at or near the species boundaries.

While the discordance between gene trees and species trees (Edwards & Beerli, 2000; Knowles & Maddison, 2002; Hey & Machado, 2003) makes interpreting the genealogical patterns observed in any single gene tree problematic (see Fig. 1), as shown by Maddison and Knowles (2006), the phylogenetic relationships of species can nevertheless be accurately inferred despite widespread incomplete lineage sorting. However, study design is critical to such inferences (Takahata, 1989; Rosenberg, 2002). This study shows that the accuracy of phylogenetic estimates differed depending on the number of loci considered, where accuracy reflects the proportion of replicates in which species relationships were estimated correctly for each species tree. A higher percentage of correct species relationship was recovered with three loci compared to one locus (average of 59.4 ± 1.81 standard error vs. 48.2 ± 1.37 standard error, respectively).

Both species divergence time and the internode distance explained a significant amount of the variance in phylogenetic accuracy across the species trees (Fig. 3). However, internode distance had a much stronger influence on phylogenetic accuracy compared to species divergence time. The ability to estimate the sister relationships of species E and F (e.g., Fig. 2) decreases dramatically with a short time interval between the divergence time of species E and F and the preceding speciation event (i.e., small internode distances). This matches theoretical expectations that demonstrate that not only the timing of species divergence, but also the interval between speciation events, strongly influences the consistency probability between species and gene trees (Takahata, 1989; Rosenberg, 2002). Even with very recent species divergence (i.e., less than $0.1 N_e$), the relatedness of species E and F was recovered accurately using genealogical information from a single locus (Fig. 2). However, if the timing of the E and F species divergence was similar to that of the preceding speciation event (i.e., short internode distance), there is a low probability of recovering the sister relationships of E and F, even with three loci (Fig. 3). ANCOVA showed that when controlling for differences in divergence time (Table 1), internode length still explained a significant amount of the variance in phylogenetic accuracy across the different species trees when both a single locus and three loci were used to estimate the species tree ($r^2 = 0.80$, $P < 0.0001$, and $r^2 = 0.68$, $P < 0.0001$, for the whole model for one locus and three loci, respectively).

The timing of divergence is not the only factor that affects phylogenetic accuracy—the relative position of taxa in the species tree (i.e., nodal position) also

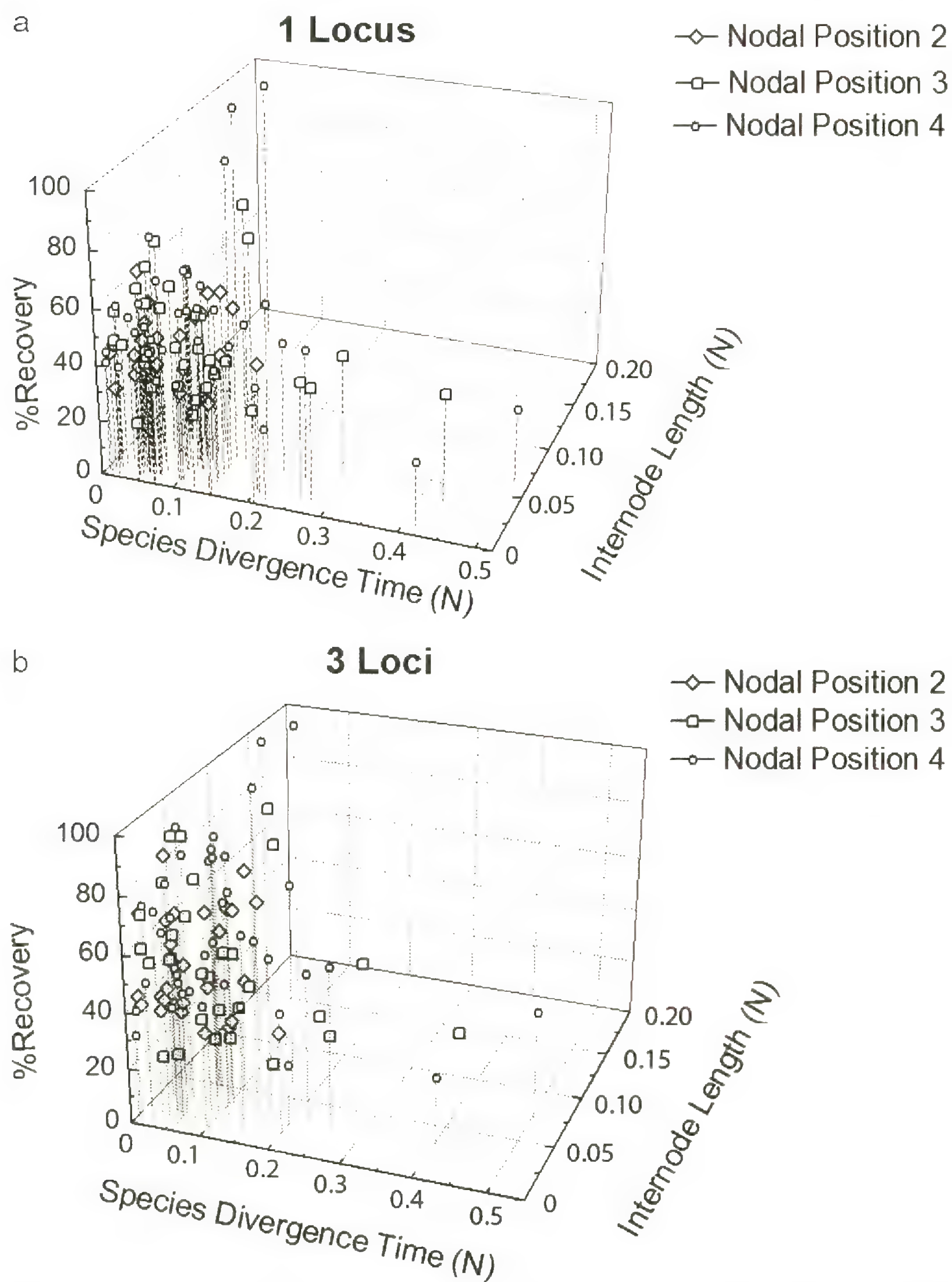


Figure 2. Phylogenetic accuracy across the range of species divergence times and internode lengths represented in the 100 simulated species trees when species relationships are estimated from one and three loci, respectively; the position of the focal taxa relative to the other species (i.e., nodal position) is marked by the different shapes.

impacted whether the sister relationships of E and F were accurately recovered (Table 1). Moreover, nodal position also influenced the effect of internode distance on phylogenetic accuracy (i.e., the interaction term in Table 1). These effects on the ability to reconstruct the sister status of E and F no doubt

reflect the probability that gene lineages from species E and F are likely to reach as deep as two, three, or more of the preceding species branchings (see also Shedlock et al., 2004). In addition to the obvious implication this observation has for the effect of species sampling on phylogenetic accuracy, it also has

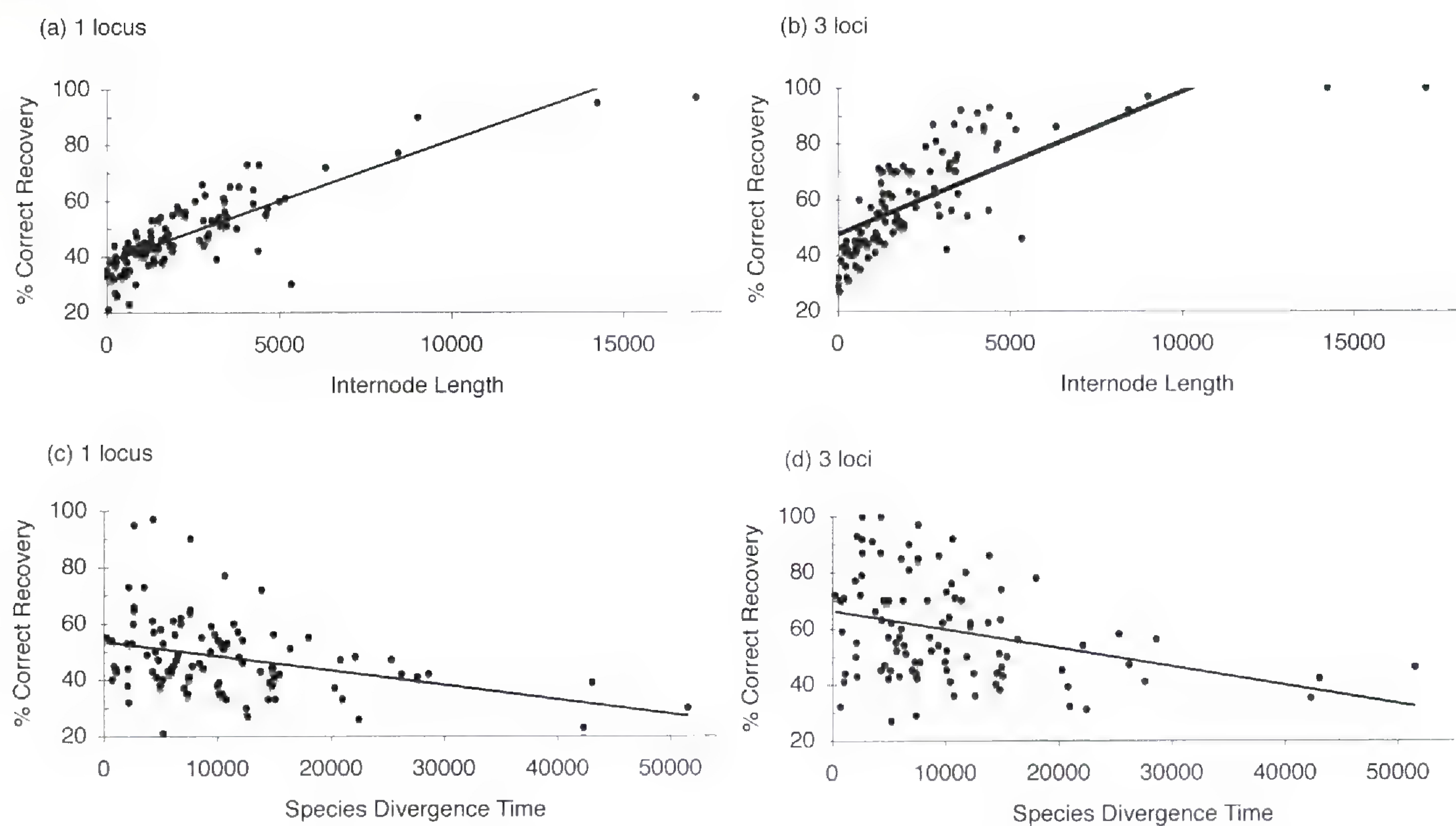


Figure 3. Phylogenetic accuracy (percent of the 100 replicate data sets for each of the 100 species trees in which the correct species relationships were recovered) is significantly influenced by both the length of the internode leading to the target taxa ($r^2 = 0.80$, $P < 0.0001$ for one locus [a] and $r^2 = 0.68$, $P < 0.0001$ for three loci [b]) and the species divergence time ($r^2 = 0.36$, $P < 0.0007$ for one locus [c] and $r^2 = 0.24$, $P < 0.001$ for three loci [d]); the effect of divergence time and internode length has been isolated (i.e., ANCOVA based on either the residuals from a regression of percent correct recovery on either internode length, or divergence time, with node position as a covariate).

important consequences for the ability to recover species relationships during evolutionary radiations. Because the species tree influences the distribution of gene genealogies (see also Degnan & Salter, 2005), the ability to recover the sister relationship of a particular pair of taxa depends on their placement relative to other species, in this case, the placement of species E and F relative to the other four species (Fig. 1). Further investigation of this effect will provide important clues into how the ability to reconstruct these relationships will change over time with the sorting of ancestral polymorphism.

The ability to recover the species relationships is unfortunately expected to diminish with the sorting of ancestral polymorphism when using an approach such as minimizing the number of deep coalescents, as suggested by the decrease in accuracy with increasing species divergence time (Fig. 3), since any signal apparent in the pattern of incomplete lineage sorting will be lost with time. While this suggests limited application of this approach for inferring relationships during evolutionary radiations that have occurred in the more distant past, it could be used as a tool to evaluate whether or not such relationships are likely

Table 1. Phylogenetic accuracy increases with increasing internode length (where, to control for differences in divergence times, the analyses were based on the residuals from a regression of accurate phylogenetic recovery on the timing of species divergence). However, this relationship depends on the shape of the species tree (i.e., the relative position of the target taxa in the species tree) as is evident from the significant effect of node position and significant interaction term in the ANCOVA.

	Degrees of freedom	Sum of squares	F-ratio	Prob > F
One locus				
Internode distance	1	7168.88	207.30	< 0.0001
Node position	2	311.52	4.50	0.0136
Internode distance $\times$ node position	2	576.41	8.33	0.0005
Three loci				
Internode distance	1	13369.94	136.38	< 0.0001
Node position	2	1512.06	7.71	0.0008
Internode distance $\times$ node position	2	2388.90	12.18	< 0.0001

to be estimated accurately from DNA sequences. This general issue of which particular evolutionary scenarios will defy accurate phylogenetic estimation has been largely overlooked despite its obvious implications for historical inference (see also Knowles & Maddison, 2002).

CONCLUSION

Despite the relatively modest sampling (i.e., one or three loci sequenced in 10 individuals per species) and challenging conditions (i.e., the radiation of six species over a time spanning just 1 *N* generations) considered here, species relationships were nonetheless accurately estimated for many of the individual species histories examined (Fig. 2). As with estimates of population genetic parameters (Felsenstein, 2006), increased sampling of loci will no doubt also increase the accuracy of estimated species relationships when the process of gene lineage coalescence is incorporated into the phylogenetic approach (Liu & Pearl, 2005; Maddison & Knowles, 2006; Carstens & Knowles, 2007). Moreover, these findings have important implications for how taxon sampling may influence the ability to recover species relationships and point to further investigation into how phylogenetic accuracy may shift as the time since speciation increases and when taxa have undergone a rapid radiation.

Literature Cited

- Arbogast, B., S. V. Edwards, J. Wakeley, P. Beerli & J. B. Slowinski. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic time scales. *Annual Rev. Ecol. Syst.* 33: 707–740.
- Avise, J. C., J. F. Shapiro, S. W. Daniel, C. F. Aquadro & R. A. Lansman. 1983. Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. *Molec. Biol. Evol.* 1: 38–56.
- Carstens, B. C. & L. L. Knowles. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanoplus* grasshoppers. *Syst. Biol.* 56: 1–12.
- Degnan, J. H. & L. A. Salter. 2005. Gene tree distributions under the coalescent process. *Evolution* 59: 24–37.
- Edwards, S. V. & P. Beerli. 2000. Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeography studies. *Evolution* 54: 1839–1854.
- , L. Liu & D. K. Pearl. 2007. High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. U.S.A.* 104: 5936–5941.
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.
- . 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Molec. Biol. Evol.* 23: 691–700.
- Hasegawa, M., H. Kishino & T. Yano. 1985. Dating of the humanape splitting by a molecular clock of mitochondrial DNA. *J. Molec. Evol.* 22: 160–174.
- Hewitt, G. M. 2004. Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans., Ser. B.* 359: 183–195.
- Hey, J. & C. A. Machado. 2003. The study of structured populations—New hope for a difficult and divided science. *Nat. Rev. Genet.* 4: 535–543.
- Hudson, R. 1990. Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7: 1–44.
- & J. A. Coyne. 2002. Mathematical consequences of the genealogical species concept. *Evolution* 56: 1557–1565.
- Jackman, T. R., A. Larson, K. de Queiroz & J. B. Losos. 1999. Phylogenetic relationships and tempo of early diversification in *Anolis* lizards. *Syst. Biol.* 48: 254–285.
- Jennings, W. B. & S. V. Edwards. 2005. Speciation history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* 59: 2033–2047.
- Kingman, J. F. C. 1982. The coalescent. *Stochastic Process Applic.* 13: 235–248.
- Knowles, L. L. 2000. Tests of Pleistocene speciation in montane grasshoppers from the sky islands of western North America (genus *Melanoplus*). *Evolution* 54: 1337–1348.
- . 2001. Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshoppers. *Molec. Ecol.* 10: 691–701.
- . 2004. The burgeoning field of statistical phylogeography. *J. Evol. Biol.* 17: 1–10.
- & W. P. Maddison. 2002. Statistical phylogeography. *Molec. Ecol.* 11: 2623–2635.
- & B. C. Carstens. 2007a. Estimating a geographically explicit model of population divergence. *Evolution* 61: 477–493.
- & ———. 2007b. Delimiting species without monophyletic gene trees. *Syst. Biol.* 56: 887–895.
- Liu, L. & D. K. Pearl. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56: 504–514.
- Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* 46: 523–536.
- & D. R. Maddison. 2004. Mesquite: A modular system for evolutionary analysis. Version 1.01. <<http://mesquiteproject.org>>, accessed 28 February 2008.
- & L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55: 21–30.
- Masta, S. E. & W. P. Maddison. 2002. Sexual selection driving diversification in jumping spiders. *Proc. Natl. Acad. Sci. U.S.A.* 99: 4442–4447.
- Mendelson, T. & K. L. Shaw. 2005. Rapid speciation in an arthropod. *Nature* 433: 375.
- Miyamoto, M. M. & W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* 44: 64–76.
- Nikaido, M. H., H. Hamilton, H. Makino, T. Sasaki, K. Takahashi, M. Goto, N. Kanda, L. A. Pastene & N. Okada. 2006. Generating single-copy nuclear gene data for a recent adaptive radiation. *Molec. Phylogen. Evol.* 39: 124–134.
- Pamilo, P. & M. Nei. 1988. Relationships between gene trees and species trees. *Molec. Biol. Evol.* 5: 568–583.
- Poe, S. & A. L. Chubb. 2004. Birds in a bush: Five genes indicate explosive evolution of avian orders. *Evolution* 58: 404–415.
- Rieseberg, L. H., T. E. Wood & E. J. Baack. 2006. The nature of plant species. *Nature* 440: 524–527.

-
- Rokas, A., B. L. Williams, N. King & S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425: 798–804.
- Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. *Theor. Populat. Biol.* 61: 225–247.
- . 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* 61: 225–247.
- Shedlock, A. M., K. Takahashi & N. Okada. 2004. SINEs of speciation: Tracking lineages with retrotransposons. *Trends Ecol. Evol.* 19: 545–553.
- Slowinski, J. B. 2001. Molecular polytomies. *Molec. Phylog. Evol.* 19: 114–120.
- Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437–460.
- Takahata, N. 1989. Gene genealogy in 3 related populations—Consistency probability between gene and population trees. *Genetics* 122: 957–966.
- & M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110: 325–344.
- Tateno, Y., M. Nei & F. Tajima. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Molec. Evol.* 18: 387–404.
- Walsh, H. E., M. G. Kidd, T. Moum & V. L. Friesen. 1999. Polytomies and the power of phylogenetic inference. *Evolution* 53: 932–937.